RADAR Project Evaluation Plan

Principal Investigators:
Daniel P. Siewiorek, Jaime G. Carbonell,
Scott E. Fahlman

RADAR Project Overview

A Grand Challenge

Build a cognitive assistant that can handle unanticipated requests and situations without reprogramming…

This requires…

- Extensive background knowledge.
- A flexible planner that can weave together plan fragments into new plans.

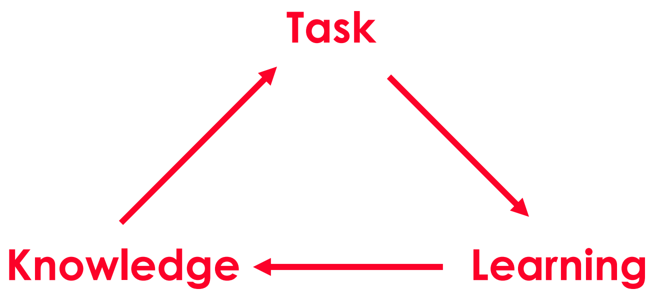
And these require…

- Learning – the only way to acquire enough knowledge, plan fragments, and strategies.

… and that improves over time.

Real-World Learning

- There has been a lot of good work on learning in simple environments. CMU is a leader.
- RADAR is a rare opportunity to study learning in a difficult RADAR-world task with a large body of complex knowledge.



Overall RADAR Goals

- Build a cognitive assistant for busy managers.
- Push learning and knowledge representation to new levels.
- Deal in some reasonable way with unexpected requests and situations.
- Crystallize common techniques into re-usable modules (“GEMs”), creating a toolkit.

Crisis Grand Challenge

Dealing with a crisis situation (sudden loss of space due to contamination) will require coordinated action by most parts of RADAR.

- Need flexible overall planning strategies.
- Intense flurry of E-mail, some urgent.
- Lots of negotiation.
- Need to plan meetings quickly.
- Need to communicate the current situation to multiple audiences.
- Space-planner pushed into area where it has little experience.

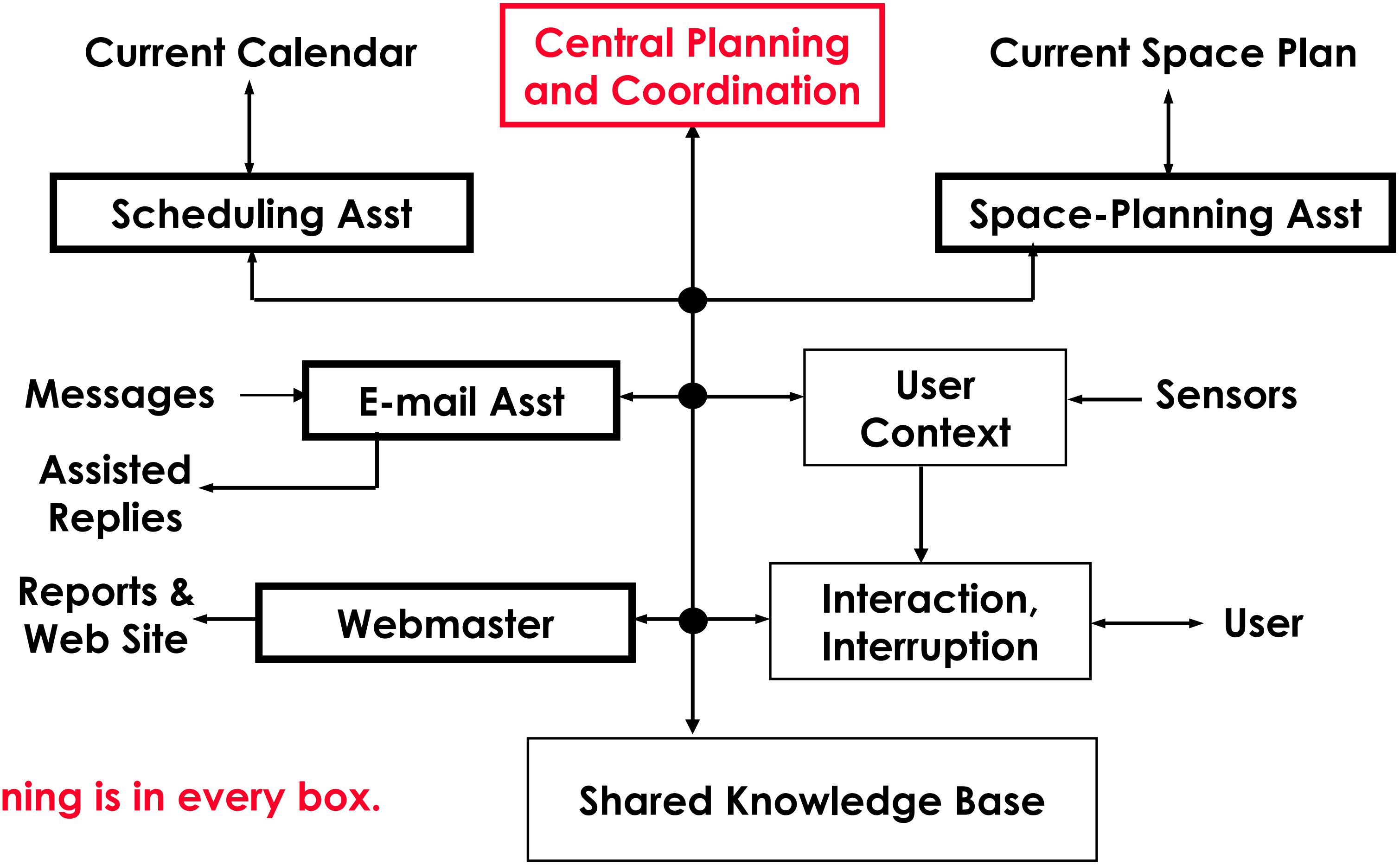
Initial Activities for RADAR

- E-Mail Assistant
  - Bill Scherlis, Eric Nyberg, Jim Herbsleb, Alex Waibel
- Virtual Information Officer (a.k.a. Webmaster)
  - Raj Reddy, Anthony Tomasic, Ravi Mosur, Alex Rudnicky
- Scheduling Assistant
  - Manuela Veloso, Steve Smith, Lori Levin
- Base-Line (Non-Crisis) Space-Planning Assistant
  - Jaime Carbonell, Eugene Fink, Bob Frederking

Plus cross-cutting modules, architecture, and user studies, evaluations.

  - Dan Siewiorek, William Cohen, Scott Fahlman, Jodi Forlizzi, Susan Fussell, David Garlan, Scott Hudson, Sara Kiesler, Bob Kraut, Tom Mitchell, Brad Myers, Brad Schmerl, Asim Smailagic, Yiming Yang, John Zimmerman

RADAR Architecture



Learning is in every box.

Testing and Evaluation

General Principles

- Independent Evaluator for years 2-5 will concentrate on evaluation of the Grand Challenge (Space Crisis) scenario, which exercises most parts of RADAR.
- Quantitative measurement of the cognitive assistant's performance is very important, but ultimate success of this program will also require qualitative breakthroughs.
- The expectation is that DARPA, the evaluator, and CMU will work together to produce an evaluation plan that evolves as the research evolves.

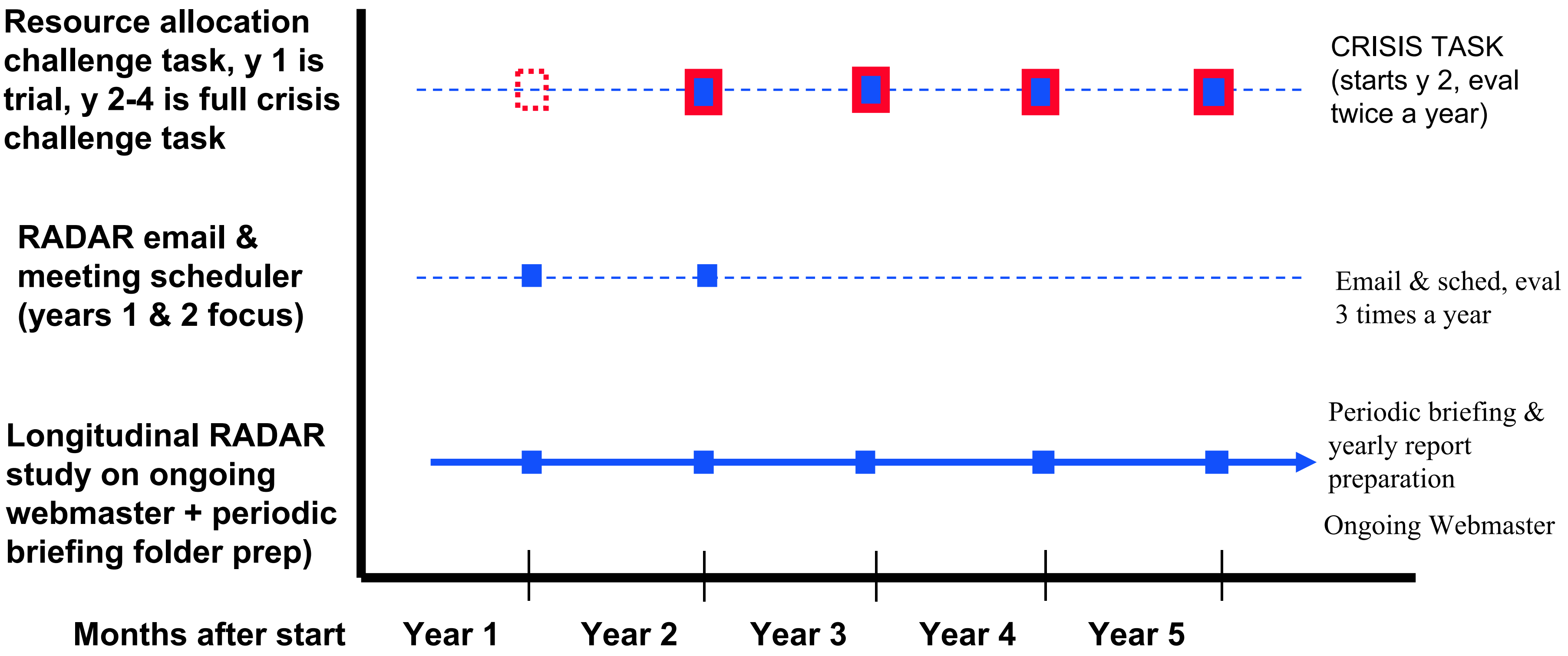
RADAR: Research Context and Goals

- User Profile - A Manager
  - Tens of communicating partners per day
  - Receives hundreds of e-mails per day including requests for meetings and information
  - Participates in many on-going, interleaved projects - tens of projects concurrently in progress
  - Documents in tens of different formats
  - Over constrained calendar
  - Initiates and coordinates meetings
  - Responds to unplanned tasks and crises
- Goal
  - Accomplish task two to four times faster than without RADAR technology
  - Accuracy and coverage up to twice as effective as without RADAR technology

Each Task has Defined Metrics

- Component - Goals identified for each component such as accuracy, coverage of knowledge contained in input
    - Efficiency - reduction in training sets
    - Quality - coverage of knowledge contained in input
- System - Goals identified for each thrust
    - Efficiency - time/effort to perform task compared to control group without RADAR technology
    - Quality - accuracy/cost of completed task compared to control group without RADAR technology
- Isolate and measure contribution of learning
- Compare to human assistant

Experimental Plan (Years 1 to 5)


Resource allocation challenge task, y 1 is trial, y 2-4 is full crisis challenge task
RADAR email & meeting scheduler (years 1 & 2 focus)
Longitudinal RADAR study on ongoing webmaster + periodic briefing folder prep)
CRISIS TASK (starts y 2, eval twice a year)
Email & sched, eval 3 times a year
Periodic briefing & yearly report preparation
Ongoing Webmaster
Months after start
Year 1
Year 2
Year 3
Year 4
Year 5

Large scale, controlled experiments with faculty and students from the project and hundreds of student subjects

RADAR Tasks: Target Performance Improvements

	Year 1	Year 2	Year 3	Year 4	Year 5
General Research Focus	Build components, simple end-to-end tasks, some use of learning	Multi-component tasks, more extensive use of learning	Coordinated operation of multiple tasks, overall planning, cross-problem learning	Dealing with surprise requests, learning applied to surprise requests, more background knowledge.	Extension to broader range of unexpected requests, improve due to user feedback & self-analysis.
E-mail and Scheduler					
Efficiency: Time to process message queue reduced by	Factor of 2	Factor of 3	-	-	-
Quality: number of short messages answered appropriately	Comparable to human	Comparable to human	-	-	-
Efficiency: Time to schedule meetings reduced	2	2+	-	-	-
Quality: Cost of meeting changes	Comparable to human	25% reduction in cost of rescheduling	-	-	-
Webmaster and Annual Report					
Efficiency: Elapsed time to posting improvement factor	1.3	2 (twice as fast)	3 (3 times as fast)	5 (up to 5 times as fast)	
Quality: Number of errors (misplaced items, duplications, broken files, broken links, etc) reduction factor	Comparable to human	1.3	1.5	2 (half the error rate)	
Efficiency: Elapsed time to assemble report, improvement factor	1.3	2 (twice as fast)	3 (3 times as fast)	5 (up to 5 times as fast)	
Quality: Completeness of annual report improvement in omission rate	Comparable to human	1.3	1.5	2 (half the non-RADAR team missing info now included)	
Space Planning Crisis Task					
Efficiency: Effort and time to converge improvement factor	-	1.2	1.5	2 (twice as fast)	Speedup equivalent to multiple human assistants
Quality:Percentage of gap between human and ideal solution closed by RADAR	-	15%	30%	50% (RADAR performs at a level half-way between human and ideal – I.e. twice as good as unaided human)	User + RADAR quality equivalent to user + multiple human assistants.

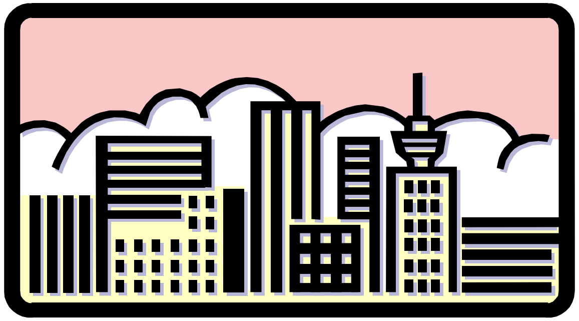
CMU Internal Evaluation

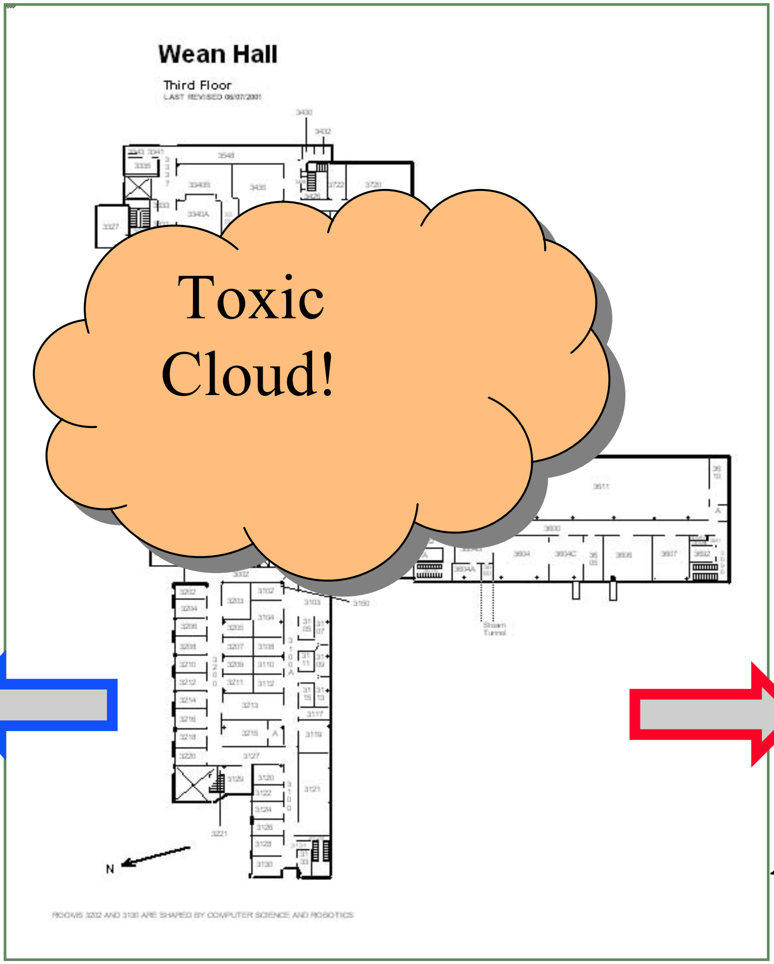
Space Crisis Grand Challenge

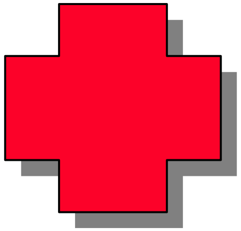
GRAND CHALLENGE: Space Planning Crisis Task

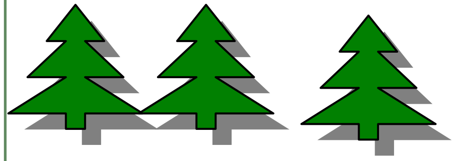
- Building housing critical functions is rendered unsafe by terrorists or natural disaster (e.g. bomb damage, anthrax spores, flood, flunk structural safety inspection, …)
- TASK: Relocate personnel & equipment into other existing facilities minimizing down time & collateral disruptions
- SWAT team of four with RADAR must plan relocation, including negotiation for space resources. Limited time of six-eight hours to complete.
  - Control team uses standard tools
  - Control team^{++} has 4 additional human assistants.
  - RADAR team has trained RADAR but no human assistants
- Metrics
  - Efficiency: Effort and time to converge on solution reduced by factor of two in year four, permitting more complex problems to be solved
  - Quality: Percentage of gap between control and ideal solution decreased by 50% for RADAR team in year four

Surprise Space Allocation → Urgent Response Challenge




Wean Hall
Third Floor
LAST REVISED 06/07/2001
Toxic Cloud!
N
ROOMS 3202 AND 3130 ARE SHARED BY COMPUTER SCIENCE AND ROBOTICS



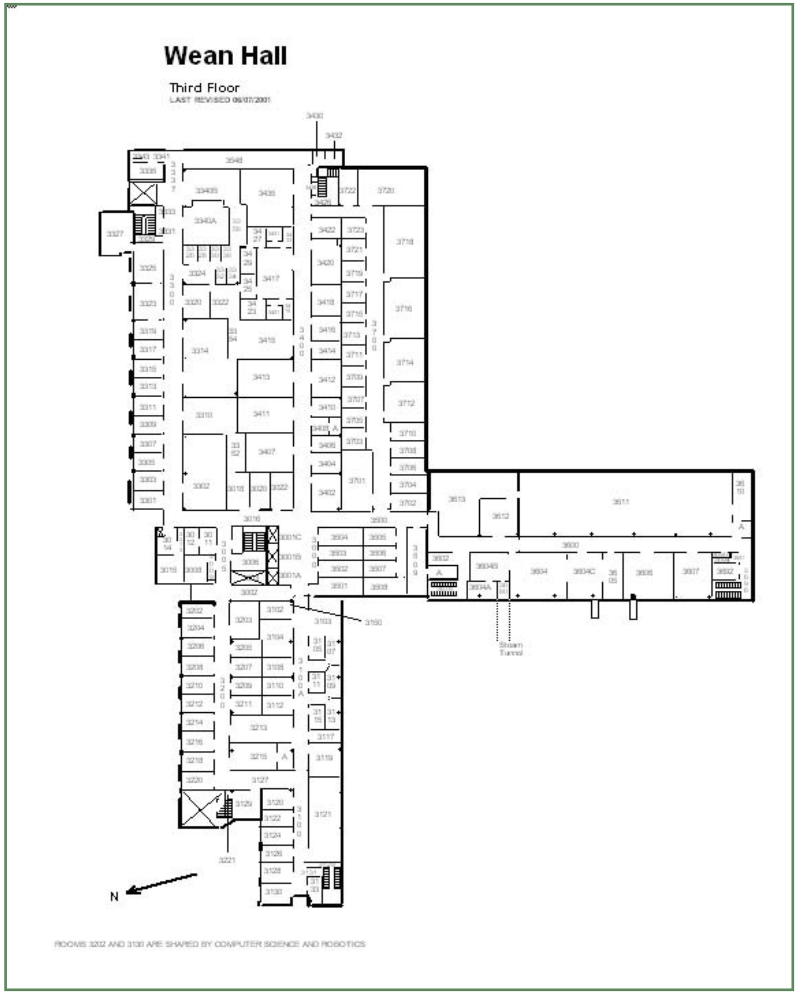


-Alloc. Solvable?

-Decomposable?

-Cope w/Surprise: (not ethernet-wired, dispersed,…)

RADAR SPACE Challenge
Why is it deep research?


Wean Hall
Third Floor
LAST REVISED 06/07/2001
N
ROOMS 3202 AND 3130 ARE SHARED BY COMPUTER SCIENCE AND ROBOTICS

- How to represent and reason about space
- How to optimize space allocation conditioned on resource, constraints, preferences, and forecasts
- How to cope with surprise (crises, degraded space, new constraints, new preferences, new utility functions, new optimization criteria, …)
- How to cope with uncertainty (partial knowledge of preferences, contingency planning based on possible exogenous events, predicting negotiation outcomes)
- How to learn what worked and why for next time: surprise → methods

Surprise Generator

- Standard, replicable, training & test sets
  - **Initial conditions** (space layout, occupants, preferences,…)
  - **Tasks** (new people to allocate, reorganization, move to new space, existing space goes away, …)
- Brand new task injection
  - **Random selection from distribution: new initial conditions or new tasks, some unsolvable**
  - **Preferences, constraints → new ones (e.g. new project)**
  - **Optimization criterion changes (e.g. new boss or mission)**
  - **Categorically-new relations** (now we have room size, spatial layout, connectivity, … → new: wet-lab enabled, faraday cage "skiff" enabled) & new constraints, preferences using these
- Metrics:
  - **Ability to solve/optimize space allocation with surprise**
  - **Analogical reasoning from earlier solutions if appropriate**

Grand Challenge Test Conditions: Escalating Space Management Crisis

Summary of Key Assumptions

Three groups composed of four people each controlling the space in one building that is filled to capacity according to current space guidelines

A fifth building closes a wing and occupants plus their equipment must be relocated into other four buildings.

Whole reallocation plan must be completed in six hours

Building closing escalates during experiment to include a second wing and finally the core

Background activity mixes with space planning negotiation

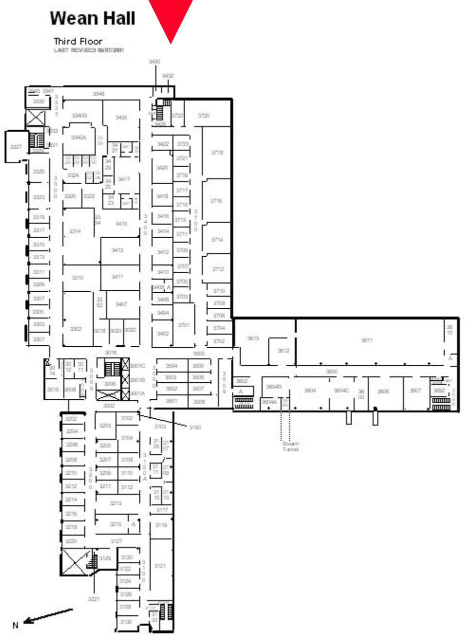
Impact on a second critical task, scheduling a series of important meetings for the following week, also measured

- E-mail and Instant Messages requesting a series of urgent meetings next week
- Background e-mail
- Non-relevant and relevant Instant Messages causing interrupts

RADAR

Control

Control ++
Each participant has a human assistant



- Collaborative Replanning
- Evaluation of Radar as Assistant

Metrics

Efficiency

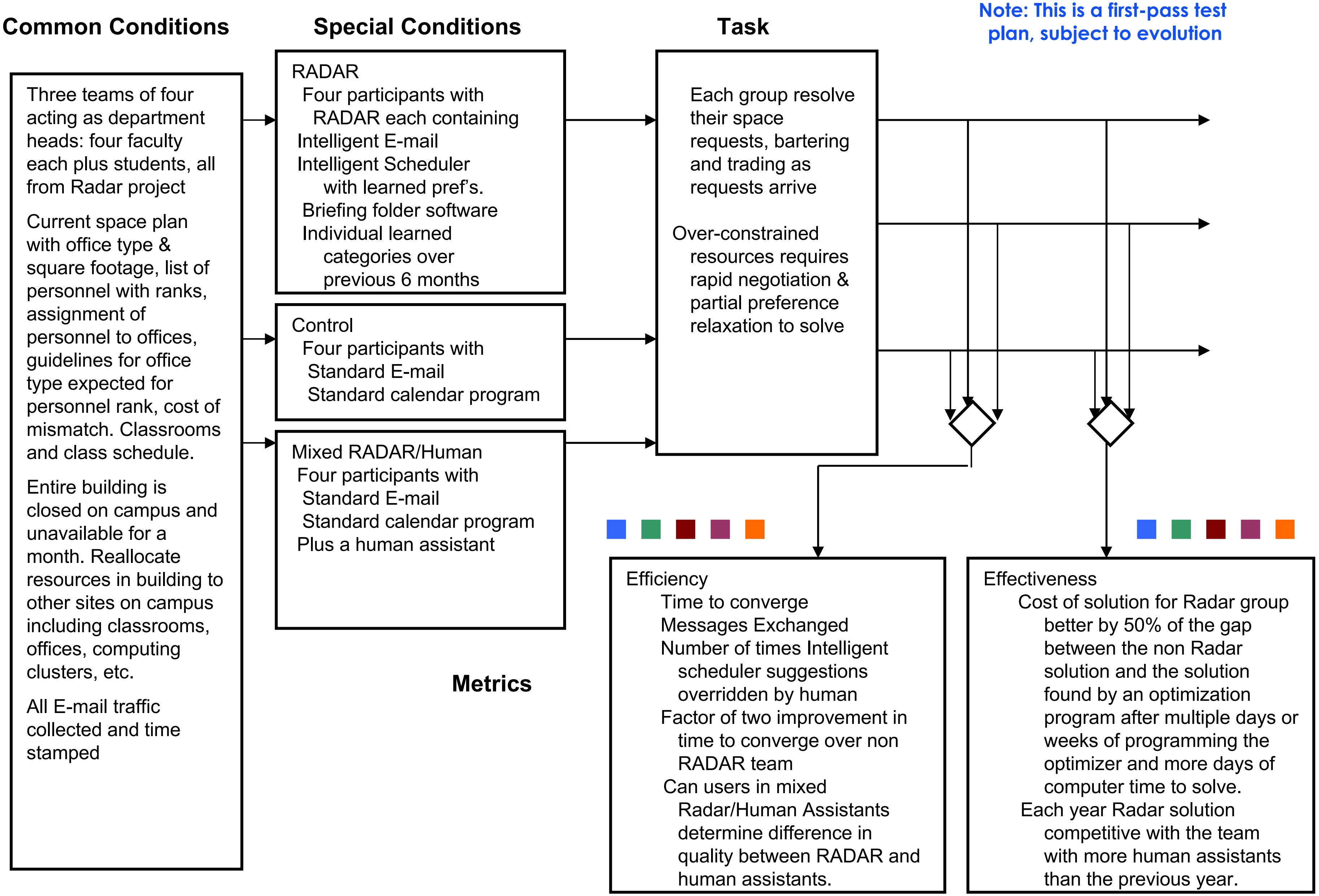
Time to converge

Factor of two over control by year four

Quality Cost of solution for Radar group better by 50% of the gap between control solution and solution found by an optimization program after days of computer time

GRAND CHALLENGE : Complex (Crisis) Task, Y 2-5

Note: This is a first-pass test plan, subject to evolution



Common Conditions

Three teams of four acting as department heads: four faculty each plus students, all from Radar project

Current space plan with office type & square footage, list of personnel with ranks, assignment of personnel to offices, guidelines for office type expected for personnel rank, cost of mismatch. Classrooms and class schedule.

Entire building is closed on campus and unavailable for a month. Reallocate resources in building to other sites on campus including classrooms, offices, computing clusters, etc.

All E-mail traffic collected and time stamped

Special Conditions

RADAR
- Four participants with RADAR each containing
- Intelligent E-mail
- Intelligent Scheduler with learned pref's.
- Briefing folder software
- Individual learned categories over previous 6 months

Control
- Four participants with
- Standard E-mail
- Standard calendar program

Mixed RADAR/Human
- Four participants with
- Standard E-mail
- Standard calendar program
- Plus a human assistant

Task

Each group resolve their space requests, bartering and trading as requests arrive

Over-constrained resources requires rapid negotiation & partial preference relaxation to solve

Metrics

Efficiency
- Time to converge
- Messages Exchanged
- Number of times Intelligent scheduler suggestions overridden by human
- Factor of two improvement in time to converge over non RADAR team
- Can users in mixed Radar/Human Assistants determine difference in quality between RADAR and human assistants.

Effectiveness
- Cost of solution for Radar group better by 50% of the gap between the non Radar solution and the solution found by an optimization program after multiple days or weeks of programming the optimizer and more days of computer time to solve.
- Each year Radar solution competitive with the team with more human assistants than the previous year.